学修番号 12890526

修士論文

英単語タイピングゲームによる
スペリング誤りの抽出と分析

立花 竜一

2015 年 9 月 9 日

首都大学東京大学院
システムデザイン研究科 情報通信システム学域

立花 竜一

審査委員：

小町 守 准教授　（主指導教員）

石川 博 教授　（副指導教員）

高間 康史 教授　（副指導教員）

英単語タイピングゲームによる スペリング誤りの抽出と分析*

立花 竜一

内容梗概

これまで英語のスペリング誤り抽出に関する多くの研究が行われてきた．それらの研究では Twitter などの web サービスからスペリング誤りの候補を抽出する，クラウドソーシングを利用した入力ログからスペリング誤りの候補を得るといった方法などでスペリング誤りを収集していた．しかし，そういった研究では抽出したスペリング誤りが何の単語のスペリング誤りかわからない，クラウドソーシングのコストがかかるといった難点があった．

そこで本研究では英単語タイピングゲームを利用することで，スペリング誤りに対応する単語が明らかであり，クラウドソーシングのコストもかからないスペリング誤りの抽出手法を提案し，実際に抽出したスペリング誤りに関する分析を行う．本研究の目的はスペリング誤り訂正に貢献することである．またスペリング誤り抽出のためにタイピングゲームを利用することで起きる結果を仮定し，仮説が正しいことを示す．

*首都大学東京大学院 システムデザイン研究科 情報通信システム学域 修士論文, 学修番号 12890526, 2015 年 9 月 9 日.

Extraction and analysis of spelling errors using English word typing game*

Ryuichi Tachibana

Abstract

There have been many studies on English spelling error extraction. Previous studies collect English spelling errors from web services such as Twitter by using edit distance or from the input log by utilizing crowdsourcing. However, in the former approach, it is not clear which word corresponds to the spelling error; and in the latter approach, it requires annotation cost for crowdsourcing.

Therefore, I propose to extract spelling errors by using an English word typing game. This game allows us to know the word corresponding to the spelling error and saves the cost of crowdsourcing. I analyze actual spelling errors extracted from the game. The purpose of this study is contribution to English spelling error correction. I assume results caused by using English word typing game for English spelling error extraction, I show that the hypothesis is correct.

*Master's Thesis, Department of Information and Communication Systems, Graduate School of System Design, Tokyo Metropolitan University, Student ID 12890526, September 9, 2015.

目次

図目次

第 1 章　はじめに

パソコンは主にキーボードを利用して操作を行うが，キーボードを利用する場合，スペリングを誤ってしまうことがある．たとえば，隣接するキーを誤って押してしまう混同や，アルファベット同士の音韻的・視覚的類似性によって間違えてしまう誤りがある．本研究ではキーボードの誤操作によって発生するタイポと誤認識によって発生する綴りの混同を合わせて**スペリング誤り**と定義する．

パソコンの普及率の向上に伴い，スペリング誤りの検出・訂正が注目され，スペリング誤りを分析してその原因を解明しようとする研究やスペリング誤り訂正に貢献する研究が行われている．たとえば，Twitter などの web サービスからスペリング誤りの候補を抽出する [1]，クラウドソーシングを利用した入力ログからスペリング誤りの候補を得る [2] などしてスペリング誤りの獲得を行う研究がある．しかしコーパスから教師なしにスペリング誤りの候補を抽出する場合何の単語のスペリング誤りかわからない，クラウドソーシングを利用する場合コストがかかるといった難点があった．

一方ゲームにおいて用いられている要素をタスクに活用することによって，ユーザがタスクにより取り組んでもらえるようにタスクの設定を工夫するゲーミフィケーションの研究が盛んになってきている [3]．教師なしに知識獲得をするのと異なり，ゲーミフィケーションでは自分でタスクを設定することができる利点がある．また，ゲーミフィケーションではユーザに対価を払うことなく，何らかのタスクにおいてのリソースを獲得できるという利点がある．

そこで本研究ではタイピングゲームを利用することで，スペリング誤りに対応した単語は何か分かっており，コストもかからないスペリング誤りの抽出手法を提案する．抽出したスペリング誤りに関する分析を行い，スペリング誤り訂正に貢献することを目的とする．タイピングゲームを利用してスペリング誤り訂正に貢献することが，一般に存在するタイピングゲームと異なっている．

本研究ではタイピングゲームを利用することから以下のような仮説を立てた．

- 本研究のタイピングゲームでは速くタイピングゲームをクリアするほどより高いスコアを獲得することができ，ユーザは急いで文字を入力するため，打鍵ミスによるスペリング誤りの割合が大きくなる．

- ユーザは入力する文字を見て入力しているため，音韻的混同が原因によるスペリング誤りの頻度が相対的に小さくなる．
- Baba らが抽出した最終出力文字列に残らない，ユーザが文字を入力中に修正するスペリング誤り [2] と最終出力文字列に残る一般的なスペリング誤りの両方を抽出するため，それぞれの誤りが混合した結果になる．

これらの仮説がスペリング誤りに関する定量的な分析において正しいかどうか抽出された結果に基づいて考察する．

スペリング誤りを抽出するためにタイピングゲームを利用することは，抽出されるスペリング誤りに影響を与えるような状況を生み出すと考えられ，その状況は以下である．

- ユーザは入力する文字を見て入力する状況になる．ユーザが入力する文字がわかるという利点があるが，入力する文字を見て入力する状況が影響したスペリング誤りの抽出になる．
- タイピングゲームにおいてユーザが急いで文字を入力する状況になる．速く文字を入力することで平常時より多くのスペリング誤りを抽出できる可能性があるが，急いで文字を入力する状況が影響したスペリング誤りの抽出になる．
- 本研究で利用したタイピングゲームでは既存のタイピングゲーム [4] と同様にユーザ自身が文字を消去する設計がない．そのため本研究での手法が既存のタイピングゲームへの応用に繋がる可能性があるが，文字の過剰誤りと置換誤りを区別することや，抽出したスペリング誤りが Baba らが抽出した最終出力文字列に残らない，ユーザが文字を入力中に修正するスペリング誤り [2] かどうか判断することができない．

本研究では荒牧ら [1] や Baba ら [2] の実験における状況と比べてこのような難点があるが，我々が抽出したスペリング誤りの結果と Baba らが抽出したスペリング誤りの結果を比較して，それぞれの結果が似たような結果であることがわかれば，タイピングゲームを利用して抽出したスペリング誤りがスペリング誤り訂正に貢献するために活用できると考えられる．

本研究の主要な貢献は以下である．

- 我々の知る限り，タイピングゲームをスペリング誤り抽出に用いた研究は本研究が初めてである．
- ユーザが入力しようとしている英単語がわかっているため，編集距離に基づくスペリング誤り抽出手法 [1] では行えないようなスペリング誤りに関しての分析が行える．
- スペリング誤り抽出のためにタイピングゲームを利用する状況での仮説が正しいことを示す．
- タイピングゲームを利用して抽出したスペリング誤りがスペリング誤り訂正に貢献するために活用できることを示す．

第 2 章　関連研究

まずスペリング誤りに関しての関連研究について説明する．Kernighan らは Noisy Channel Model を使って単語が与えられたときの訂正候補の確率をモデル化することでスペリング訂正を行った [5]．この研究ではスペリング訂正モデルは 1 文字同士の挿入，削除，置換，転置といった編集距離の値に基づいて計算された．また Brill らは Kernighan らと同様に Noisy Channel Model によるスペリング誤りの訂正を行ったが，訂正候補の確率をモデル化するときに一文字だけでなく，文字列の編集距離を考慮することでスペリング誤りの訂正の精度を向上させた [6]．Ahmad らはウェブ検索クエリログからスペリングを訂正するためのモデルを自動的に学習した [7]．この研究では EM アルゴリズムを手法とすることで，正解データを必要とせずに重み付き編集距離を学習した．重み付き編集距離は a と e は誤りやすいが，a と l は誤りにくいといった文字同士の誤りやすさに関する情報を反映するために利用される．これらの研究ではスペリング誤りの訂正のために研究を行っているが，本研究ではスペリング誤りの訂正は行わず，スペリング誤りの特徴の分析を行っている．

Cook による研究 [8] では英語を母語とする第一言語話者と母語としない第二言語話者の英語のスペリング誤りを比較して分析を行っている．本研究では英語を母語としないユーザのスペリング誤りを分析している点が共通しているが，Cook はスペリング誤りのデータを学生がテストや宿題で書いたものなどから抽出していたのに対し，本研究ではタイピングゲームを用いてスペリング誤りを抽出する．

荒牧らは Twitter のクロールデータを利用することで英単語のスペリング誤りの抽出を行い，スペリング誤りの原因を分析した [1]．またスペリング誤りとスペリング誤りでないものを判別する学習器を構築し，実験を行うことで分析結果の検証を行った．クロールしたデータにおいてスペリング誤り候補を決定するために，英単語から編集距離が 1 であるものを収集した．本研究における荒牧らの研究との共通点は，キー配置によるスペリング誤りや単語内のスペリング誤りの位置といった観点でスペリング誤りの分析を行うことである．また荒牧らの研究では正解の単語を推測するために編集距離を利用していたが，本研究では正解の単語を推測する必要がない手法を提案し，本研究で抽出したスペリング誤りの結果と荒牧らが抽出し

たスペリング誤りの結果を比較することで，タイピングゲームを利用する状況における仮定が正しいことを考察する．

Baba らはスペル訂正エンジンの精度向上を目的としてクラウドソーシングの 1 つである Amazon Mechanical Turk を利用して，ある画像が何を描写しているかユーザに英文を入力させユーザのキーストロークを抽出し，最終出力文字列に残らない，ユーザが文字を入力中に修正するスペリング誤り候補を抽出する手法を実装し，その結果に対して分析を行った [2]．修正前文字列と修正後文字列の編集距離が 2 以下であるものを抽出し，それらを比較してスペリング誤り候補を抽出することで分析を行った．Baba らが抽出したスペリング誤りでは置換誤りの割合が多く一般的な英語のスペリング誤りでは脱落誤りの割合が多いため，ユーザにとって置換誤りは気づきやすく入力中に修正するが，脱落誤りは気づきにくく最終出力文字列に残りやすいことを報告している．この研究では Amazon Mechanical Turk を利用するためコストがかかるのに対し，本研究ではコストのかからないタイピングゲームを用いたスペリング誤りの抽出手法を提案する．

またユーザがゲームを行うことを通して言語資源を得るという研究がある．Kumaran らはある句に対して同じ意味の句を得るためにユーザにゲームを行わさせた [9]．そのゲームはあるユーザがある句に関しての絵を描き，その絵を他のユーザに見せてその絵が何の句を示しているのか答えてもらうというものであった．Kumaran らの研究と本研究ではゲームを利用して言語資源を得るという手法は同様であるが，Kumaran らは同じ意味の句対，本研究ではスペリング誤りを言語資源として得ようとすることがそれぞれ異なる．Vannella らは意味的な知識の検証と拡張のために 2 つのゲームを作成し，ユーザに行わさせた [10]．2 つのゲームはそれぞれシューティングゲームとロールプレイングゲームを模したものであり，それらのゲームを行うことで概念同士の，または概念と画像の関係のアノテーションを行わせた．ユーザがゲームを通して行ったアノテーションは，クラウドソーシングにおいて労働者が報酬を受け取って作成したアノテーションと比較しても高い品質のものであった．Venhuizen らは Wordrobe と呼ばれるゲームをユーザに行わせることで語義のラベル付けを行った [11]．ゲームは語義に関する複数の選択がある質問の集合で構成されていて，複数のユーザはそれに答えて，他のユーザと答えが一致しているかどうかによってユーザが獲得する得点が決まる．Kumaran らや

Venhuizen らの研究と本研究との共通点はゲームを用いて言語資源獲得を行う点であるが，相違点は得ることができる言語資源がそれぞれの研究において異なる点である．

第 3 章　タイピングゲームによるスペリング誤り抽出

本研究ではユーザがタイピングゲーム*を行うことで，ユーザがタイプしたアルファベットの文字列を抽出し，その文字列に対応した英単語と比較を行うことでスペリング誤りの獲得を行う．ユーザが入力する文字列にはスペリング誤りが含まれているため，ユーザがタイプしたアルファベットの文字列と英単語を比較することでスペリング誤りの分析が可能になる．

ユーザが入力したアルファベットの文字列とその文字列に対応した英単語の比較は，タイピングゲームにおいて間違いだと判定される文字を抽出するように比較が行われる．例えば表 3.1 ではユーザが入力した文字列とその文字列に対応した英単語の比較例を示している．入力すべき英単語は belief だったが，ユーザが実際に入力した文字列は beleief であったとき，belief の 4 文字目にあたる i を入力すべきときにアルファベット e を入力しているので，アルファベット e をスペリング誤りとして抽出を行う．スペリング誤りのアルファベット e と，ユーザが入力しようとしていた belief の 4 文字目にあたる i と，その前後の文字をスペリング誤りの原因の分析のためにそれぞれ抽出する．

ユーザが入力した文字列において複数のスペリング誤りが存在し，それらがタイピングゲームにおいて正解だと判定される文字または文字列で区切られている場合はそれぞれのスペリング誤りは別々のスペリング誤りとして扱うものとする．表 3.2 では 1 つの文字列において複数のスペリング誤りが抽出される例を示しており，ユーザが入力した文字列 ingcresase において 3 文字目の g と 7 文字目の s をスペリング誤りとして抽出している．3 文字目の g をスペリング誤りとして扱ったあと，ingcresase において 4 文字目から 6 文字目 cre は文字列に対応した英単語 increase の 3 文字目から 5 文字目と対応して，タイピングゲームにおいて正解だと判定される．そのあとユーザは英単語 increase の 6 文字目の文字 a を入力する必要があるが，ユーザは文字列 ingcresase の 7 文字目の s を入力しているので，結果 g と s がスペリング誤りとして抽出される．

タイピングゲームにおいてユーザがスペリング誤りだと判定される文字を連続で入力する場合もあるが，その場合連続した文字列がタイピングゲームにおいて入力

*http://cl.sd.tmu.ac.jp/~ryu/typing_game.html

表 3.1　ユーザが入力した文字列と文字列に対応した英単語の比較とスペリング誤りの抽出例

入力した文字列	対応した英単語	スペリング誤り	1 つ前の文字	入力すべき文字	1 つ後の文字
beleief	belief	e	l	i	e

表 3.2　スペリング誤りと判定される文字列，されない文字列と複数のスペリング誤りが抽出される文字列の例

ユーザが入力した文字列	文字列に対応した英単語	誤りの文字	誤りの文字列
ingcresase	increase	g と s	
inlyonly	only	i	
biyyt	bit	y と y	yy

すべき文字以降の英単語の部分文字列であれば，その文字列はスペリング誤りとは判定せずに分析を行う．表 3.2 にはスペリング誤りと判定されない場合の文字や文字列の例が示されており，ユーザが入力した文字列は inlyonly で，文字列に対応した英単語は only なので，タイピングゲームにおいて間違いだと判定される文字はそれぞれ i と n と l と y であるが，inly において文字列 nly はそれぞれ入力すべき文字 o 以降の部分文字列なので，文字列 nly はそれぞれスペリング誤りとして抽出をしないものとする．

また表 3.2 にはスペリング誤りと判定される文字列の例が示されており，ユーザが入力した文字列 biyyt から抽出されるスペリング誤りの文字として y が 2 文字抽出され，入力すべき文字は t とする．

第 4 章　タイピングゲームの設計と実装

この節ではタイピングゲームの設計と実装に関して説明する．

4.1　タイピングゲームの設計

タイピングゲームは図 4.1 のように表示され，ユーザは表示されている英単語を入力していく．ユーザが任意のキーを入力したときにゲームが開始され，それと同時にタイピングゲームにかかる時間の計測が始まる．英単語は既に入力し終わっている文字を灰色，スペリングを誤った文字を赤色，まだ入力していない文字を黒で示している．1 つの英単語を入力し終わるごとに新たな単語が表示される．100 単語タイプし終わるとゲームクリアとなる．スコアはスペリングを誤るたびに 10 点減点され，100 問タイプし終わった時間が速いほど加点される．ユーザがスペリングを誤る瞬間にスコアが減点される．プレイに応じてユーザの現在のタイムやスコアも表示される．ユーザがタイピングゲームを行った中で最も高いスコアはハイスコアとして表示される．またスペリングを誤ったときや英単語を 1 単語入力し終わる度に効果音が鳴る．

4.2　タイピングゲームの実装

「初めて学ぶ人のための JavaScript 入門:タイピングゲームを作る」*を参考にしてタイピングゲームの実装を行った．サイトではユーザのキーコードの入力を受けてその入力が表示されているアルファベットと一致しているかどうかを判定し，ユーザがアルファベットを 200 文字正しく入力するまでの時間を計るアルゴリズムを示すソースコードが書かれている．本研究ではそのコードを以下のように改変・追加することで，研究のためのタイピングゲームの実装を行った．

まずユーザに入力させていたアルファベットを英単語に変更した．次にユーザが正しく英単語を入力したときは正解を示す効果音，スペリング誤りをしたときは不

*http://www.pori2.net/js/key/3.html

正解を示す効果音が鳴る機能を実装した．またユーザが英単語をタイプしているとき，スペリングを誤った文字を赤く，既に入力し終わっている文字を灰色で表示させるように実装を行った．

またタイピングゲームにゲーム的な要素を取り入れるため，ユーザがタイピングゲームを行うのにかかった時間とともにタイピングの正確さや速さを元に算出したスコアを表示させるようにした．スコアはゲームが終了するまでの時間が短いほど，ユーザがスペリング誤りをした回数が少ないほど高くなるように設定した．

最後にユーザが英単語を 1 単語入力し終わる度にユーザが実際に入力したアルファベットの文字列とそのときタイピングゲーム上で表示していた英単語，ユーザが英単語を入力するのにかかった時間をサーバに送るように実装を行った．

本研究ではタイピングを行ったときに鳴る効果音はウェブサイト[†]で公開されているものを使用している．使用に関しては Creative Commons ライセンス[‡]に基づいて素材の改変が可能なものを使用している．

タイピングゲームでタイプする英単語としてベーシック英語 [12] を用いた．これは言語学者のチャールズ・ケイ・オグデンによって考案された英語の体系で，基本単語 850 語リストは英語の初級者向け語彙として使われており，このうち冠詞 a などを取り除いた 842 語をタイピングゲームの問題にすることから，タイピングゲームを行うことによって英語学習にも役立つ可能性があると考えられる．ユーザが打ち込むベーシック英語の英単語とその訳語はウェブサイト[§]で公開されているものを使用している．

[†] http://musicisvfr.com/free/se/quiz01.html

[‡] http://creativecommons.org/licenses/by/4.0/deed.ja

[§] http://www.catch.jp/wiki/index.php?english%2F800_Basic_English

図 4.1　タイピングゲームのスクリーンショット

第 5 章　スペリング誤りの結果と分析

情報工学系の日本人大学院生 7 人にタイピングゲームを行わせた．ユーザが英単語を入力した回数が合計で 4,724 回あり，そのなかでユーザが英単語をタイプするとき 1 度はスペリング誤りを起こしている場合が 712 回存在した．アルファベットの文字列と英単語の編集距離は最大で 10 の差が存在した．ユーザが入力した 1 つの文字列には複数のスペリング誤りが存在する場合があり，そういった場合を考慮するとスペリング誤りの合計は 859 個存在した．それらのスペリング誤りに対して分析を行う．スペリング誤りの分析では定量的な分析と定性的な分析を行う．定量的な分析ではタイピングゲームを利用する状況での仮説が正しいかどうか考察する．また情報工学系の日本人大学院生 7 人のそれぞれのユーザにおいて実験結果の個人差があることが考えられるが，個人差に関する比較は行っていない．

5.1　誤りに関する定量的な分析

抽出したスペリング誤りの文字を頻度順に並べ，そのスペリング誤りの文字に対して，入力すべき文字と入力すべき文字の前後の文字の最も頻度の高かった文字を表 5.1 に示す．頻度が等しい場合は複数の文字を示してある．入力すべき文字が文字列に対応した英単語の語頭や語末の文字であった場合，表 5.1 には記号の_を表示している．またそれぞれのアルファベットの誤りに対して，全ての入力すべき文字と入力すべき文字の前後の文字の頻度を付録 A において示す．表 5.1 の結果から a, i, u, e, o といった母音がスペリング誤りの文字の頻度の上位 10 件に含まれており，これは単語を入力する上で母音を入力する頻度が多いからだと考えられる．またタイピングゲームにおいて入力すべき文字の 1 つ後の文字を間違えて入力してしまう脱落誤りの割合が 34.2%，入力すべき文字の 1 つ前の文字を間違えて入力してしまう場合の誤りの割合が 4.9% となっており，入力すべき文字を繰り返して入力してしまうことよりも，入力すべき文字を 1 文字飛ばして入力している場合がよく起きていることがわかる．*これはタイピングゲームではハイスコアを競って急いで

*本研究では文字の挿入誤りと置換誤りの区別ができない

表 5.1　スペリング誤りの文字と頻度（入力すべき文字が語頭または語末の場合は_を示す）

スペリング誤りの文字	頻度	1 つ前の文字	頻度	入力すべき文字	頻度	1 つ後の文字	頻度
e	93	r	14	r	17	e	36
r	73	e	12	e	19	r	26
s	67	_	21	c	17	e	15
i	60	_	9	o	18	i	30
o	59	_	19	p	12	o	24
n	56	i	24	o	22	n	40
a	54	_	14	e	10	a	22
t	54	g	9	h, r	8	t	21
u	35	o	8	y	11	u	14
d	34	_	7	s	10	_	8
g	34	n	11	t	8	_	17
l	31	m	6	p	9	l	21
h	25	_	7	f	5	h, o	5
v	25	_	9	b	12	e	11
p	24	l, t	5	o	16	p	5
y	23	a	4	t	9	y	9
c	20	e, i	4	v	6	c, e	6
k	20	a	8	l	14	_	9
w	20	_	8	e	14	w	5
b	15	i	4	v	9	e	9
m	14	_	4	n	4	m	3
f	11	_	5	r	3	_	3
x	3	e, n, o	1	c	2	e	2
j	3	c	2	k	2	_	2
z	2	e, _	1	a, v	1	i, m	1
q	1	p	1	e	1	c	1

文字を入力しようとすることが原因だと考えられる．Baba らの報告 [2] では一般的な英語のスペリング誤りの脱落誤りの割合が 45% 前後，Baba らが抽出した英語のスペリング誤りの脱落誤りの割合が 22% 前後あり，Baba らの報告と，本研究では Baba らが抽出した最終出力文字列に残らないスペリング誤り [2] と最終出力文字列に残る一般的なスペリング誤りの両方を抽出していることを考慮すると，我々

が抽出したスペリング誤りの結果と Baba らが抽出したスペリング誤りの結果はそれぞれ似たような結果であると考えられる．

以下ではキーボードのキー配置によるスペリング誤り，音韻的混同，単語内のスペリング誤りの位置に対するスペリング誤りの観点で分析を行う．

5.1.1 キー配置によるスペリング誤り

打鍵するときにキーボードのキー配置が近いことからスペリング誤りをしてしまうような例がみられた．表 5.1 においてスペリング誤りの文字 e や r は互いに r と e によく打ち間違えている場合が該当する．他にも o と p，b と v のようなスペリング誤りに対してこの場合が原因の一つとして考えられる．

また荒牧らによる研究 [1] において e と a，e と i の文字の置きかわりがよく発生することが報告されているが，表 5.1 や表 5.2 を見ると e と a，e と i のスペリング誤りの頻度は e と r の頻度と比べて少なく，キー配置が原因で起こると考えられる e と r のスペリング誤りの方が今回の結果では多く見られた．これは今回の研究ではユーザ自身が修正してしまうような Baba らが抽出したスペリング誤り [2] も抽出していることが要因として考えられ，このことは Baba らが抽出した最終出力文字列に残らないスペリング誤りと最終出力文字列に残る一般的なスペリング誤りの両方の誤りが混合した結果になるという仮説を裏付けるものだと考えられる．

また隣り合ったキー同士のスペリング誤りのペアの割合は 43.3% になったことからも，キー配置による誤りが多いと考えられ，この結果は本研究のタイピングゲームではユーザは急いで文字を入力するため，打鍵ミスによるスペリング誤りの割合が大きくなるという仮説を裏付ける結果だと考えられる．

5.1.2 音韻的混同

表 5.1 で b と v は相互にスペリング誤りを起こす頻度が高い文字のペアとなっている．b と v は打鍵誤りが原因のスペリング誤りの一つであるが，これは日本人特有の音韻的混同が原因とも考えられる．r と l の置きかわりも日本人特有の音韻的混同が原因だと考えられるが，今回の結果では r と l のスペリング誤りはスペリン

表 5.2　e と i のスペリング誤りの文字と頻度

スペリング誤りの文字	入力すべき文字	頻度
e	i	10
e	a	11
a	e	10
i	e	4

表 5.3　単語内のスペリング誤りの位置の割合

長さ	頻度	1 文字	2 文字	3 文字	4 文字	5 文字	6 文字	7 文字	8 文字	9 文字
3 文字	49	0.41	0.14	0.45						
4 文字	167	0.246	0.210	0.275	0.269					
5 文字	196	0.194	0.173	0.179	0.276	0.179				
6 文字	149	0.128	0.087	0.195	0.248	0.195	0.148			
7 文字	120	0.142	0.083	0.150	0.200	0.133	0.142	0.150		
8 文字	76	0.04	0.09	0.11	0.16	0.11	0.18	0.22	0.09	
9 文字	58	0.05	0.10	0.16	0.09	0.12	0.05	0.12	0.22	0.09

グ誤りの文字が r で入力すべき文字が l のときの頻度が 6 件，スペリング誤りの文字が l で入力すべき文字が r のときの頻度が 0 件であったので，タイピングゲームにおけるスペリング誤りでは音韻的混同よりキー配置が原因であると考えられる．この結果はユーザは入力する文字を見て入力しているため，音韻的混同が原因によるスペリング誤りの頻度が相対的に小さくなるという仮説を裏付ける結果だと考えられる．

5.1.3　単語内のスペリング誤りの位置

タイピングゲームにおいて表示される英単語の文字に対して，どの位置にある文字に対してスペリング誤りが起きるかに関して分析を行う．この観点における分析は荒牧ら [1] や Baba ら [2] も行っており，荒牧らは単語の語頭や語末でのスペリング誤りの頻度は語の中頃より少なくなっていることを報告した．また Baba らは語頭での文字の脱落誤りや語末での文字の過剰誤りはユーザが気がつきやすく修正されるが，語の中頃での文字の脱落誤りや挿入誤りは気づきにくく，スペリング誤り

として残る傾向があることを報告した．

表 5.3 をみると語の中頃のスペリング誤りの割合が明確に多いとは言えない．これは本研究で抽出したスペリング誤りには Baba らが抽出したユーザ自身が修正してしまうようなスペリング誤り [2] も含まれていることが原因であると考えられる．この結果は Baba らが抽出した最終出力文字列に残らないスペリング誤りと最終出力文字列に残る一般的なスペリング誤りの両方の誤りが混合した結果になるという仮説を裏付ける結果だと考えられる．

5.2 誤りに関する定性的な分析

以下では単語同士の見間違い，同じ文字が連続している文字列に対するスペリング誤りの観点で分析を行う．この節でのスペリング誤りは目視で確認した．

5.2.1 単語の見間違い

ユーザが綴りの似ている単語同士を見間違えていると考えられるスペリング誤りが存在した．表 5.4 で文字列に対応した英単語は though だが，一方でユーザが入力した文字列は through となっている．

また綴りが似ているだけでなく，発音が単語の見間違いに影響していると考えられるスペリング誤りが抽出された．表 5.4 でそのような例を示しており，文字列に対応した英単語は here だが，一方でユーザが入力した文字列は hearhe となっていて，これはユーザが英単語 here に対して英単語 hear を入力したのではないかと考えられる．また here と hear は発音的には同じなので，タイピングゲームにおいて here を入力するという状況にも関わらず，英単語同士の発音の近さが影響してこういった誤りが起こったのではないかと考えられる．

5.2.2 同じスペリング誤りが連続している場合

タイピングゲームにおいて同じ文字が連続しているような英単語が表示されたとき，その文字列に対して同じスペリング誤りを繰り返してしまうといった事例が見

表 5.4　単語同士を見間違えた例と同じスペリング誤りを繰り返した場合の例

ユーザが入力した文字列	文字列に対応した英単語
through	though
hearhe	here
stiggff	stiff
sewwteet	sweet

られた．表 5.4 では英単語 stiff に対してユーザが文字列 stiggff を入力した場合を示しており，これは文字 f に対して文字 g を入力してしまっている．このスペリング誤りはキーボードの配置による打鍵誤りによるものだと考えられ，同じ文字が連続している文字列に対してスペリングを誤った場合，同様のスペリング誤りを繰り返してしまう傾向にあると考えられる．

また文字と文字列のアルファベットが入れ替わる場合がある．表 5.4 では英単語 sweet に対してユーザが文字列 sewwteet を入力していて，この場合ユーザは文字 w を入れようとしているときに文字 e を，文字列 ee を入れようとしているときに文字列 ww を入力してしまっていると考えられる．この事例から文字と文字同士だけでなく，文字列と文字列同士のアルファベットの入れ替わりが起こりうることを示している．

第 6 章　おわりに

本研究ではユーザがタイピングゲームを行うことでスペリング誤りを抽出する手法を提案した．ユーザ 7 名にタイピングゲームのプレイを依頼し，4,724 回の英単語タイピングログを収集し，859 個のスペリング誤りを抽出した．抽出されたスペリング誤りに対して分析を行うことで，タイピングゲームのような通常より素早くタイピングを行ったり，文字を書き写すような状況では，キー配置によって起こる誤りや入力すべき文字を飛ばしてしまう誤りがよく起きることがわかった．

本研究において，

- 本研究のタイピングゲームでは速くタイピングゲームをクリアするほどより高いスコアを獲得することができ，ユーザは急いで文字を入力するため，打鍵ミスによるスペリング誤りの割合が大きくなる．
- ユーザは入力する文字を見て入力しているため，音韻的混同が原因によるスペリング誤りの頻度が相対的に小さくなる．
- Baba らが抽出した最終出力文字列に残らない，ユーザが文字を入力中に修正するスペリング誤り [2] と最終出力文字列に残る一般的なスペリング誤りの両方を抽出するため，それぞれの誤りが混合した結果になる．

という仮説を立てスペリング誤りに関する定量的な分析を行ったが，その仮説が正しいということがわかった．また我々が抽出したスペリング誤りの結果と Baba らが抽出したスペリング誤りの結果を比較して，それぞれの結果が似たような結果であることがわかったため，タイピングゲームを利用して抽出したスペリング誤りがスペリング誤り訂正に貢献するために活用できると考えられる．

しかし本研究での実験設定ではユーザが入力しようとしている英単語はわかっているが，文字の挿入誤りと置換誤りの区別ができない．区別を可能にするためには，Baba らの研究 [2] のようにバックスペースを入力させて文字を消去させる必要があると考えられる．また本研究での実験設定ではわからないが，ユーザによってスペリング誤りの頻度や傾向が変わっていることが考えられる．そのためタイピングゲームを行うときにユーザにアカウントを登録させてユーザそれぞれを区別できるようにすることで，ユーザごとの誤り訂正モデルの構築が可能であると考えら

れる．

今後はユーザに対して教育的であるようなデータセットの利用，実験設定を検討したい．たとえば英語のリスニングゲームとしてタイピングゲームを発展させることで，新たなスペリング誤りの傾向の発見や研究において有用なログの作成，ユーザにとって有益となるようなコンテンツの構築に繋げていきたい．またタイピングゲームをアプリケーションに繋げることで大規模なスペリング誤りの抽出を可能にしたい．

謝辞

本論文の作成や研究生活など終始適切な助言を賜り，ご指導頂いた指導教員の小町守准教授に感謝致します.

本論文の副査を快く引き受けて下さった石川博教授，高間康史教授に感謝致します.

小町研究室のメンバーとは研究活動において日常的に様々な議論をして多くの知識や示唆を頂きました．また研究以外の活動を共に行うことで精神的にも支えられました。ありがとうございました。

本実験を行うにあたりタイピングゲームをプレイして頂いたユーザ 7 名に対して深く感謝いたします.

参考文献

[1] 荒牧英治, 宇野良子, 岡 瑞起:TYPO Writer:ヒトはどのように打ち間違えるのか?, 言語処理学会第 16 回年次大会予稿集, pp. 966-969 (2010).

[2] Baba, Y. and Suzuki, H.: How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 373-377 (2012).

[4] インターネットでタイピング練習 e-typing: http://www.e-typing.ne.jp/english/.

[3] Deterding, S., Dixon, D., Khaled, R. and Nacke, L.:From game design elements to gamefulness: Defining “” Gamification “” , Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek ’ 11, pp. 9-15 (2011).

[5] Kernighan, M. D., Church, K. W. and Gale, W. A.: A spelling correction program based on a noisy channel model, Proceedings of the 13th Conference on Computational Linguistics-Volume 2, pp. 205-210 (1990).

[6] Brill, E. and Moore, R. C.: An improved error model for noisy channel spelling correction, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 286-293 (2000).

[7] Ahmad, F. and Kondrak, G.: Learning a spelling error model from search query logs, Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 955-962 (2005).

[8] Cook, V. J.: L2 users and English spelling, Journal of Multilingual and Multicultural Development, Vol. 18, No. 6, pp. 474-488 (1997).

[9] Kumaran, A., Densmore, M. and Kumar, S.: Online gaming for crowdsourcing phrase-equivalents, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1238-1247 (2014).

[10] Vannella, D., Jurgens, D., Scarfini, D., Toscani, D. and Navigli, R.: Vali-

dating and extending semantic knowledge bases using video games with a purpose, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1294-1304 (2014).

[11] Venhuizen, N., Basile, V., Evang, K. and Bos, J.: Gamification for word sense labeling, Proceedings of the 10th International Conference on Computational Semantics (IWCS-2013), pp. 397-403 (2013).

[12] Ogden, C. K.: Basic English: A general introduction with rules and grammar, London: Paul Treber & Co., Ltd. (1930).

発表リスト

[NL222] 立花竜一, 小町守:英単語タイピングゲームによるスペリング誤りの抽出と分析, 研究報告自然言語処理 (NL), 2015-NL-222(10), 1-7 (2015-07-08).

付録 A　入力すべき文字とその文字の前後の文字の頻度

アルファベットの誤りに対して，入力すべき文字と入力すべき文字の前後の文字の頻度をそれぞれの表で示している．

表 A.1　入力すべき文字の頻度を示した表

誤り	入力すべき文字の頻度																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	1	3	0	10	0	1	0	8	0	0	4	1	0	5	0	1	6	7	6	0	0	0	0	1	0
b	0	0	0	1	0	0	0	1	0	0	0	1	0	2	0	0	0	1	0	0	0	9	0	0	0	0
c	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4	3	0	6	0	5	0	0
d	1	2	1	0	5	4	2	1	0	1	0	3	0	0	0	0	0	1	10	1	0	0	0	2	0	0
e	11	0	4	9	0	1	0	2	10	0	0	2	1	5	4	2	0	17	2	8	3	0	11	0	1	0
f	0	0	0	1	1	0	2	2	0	0	0	1	0	0	1	0	0	3	0	0	0	0	0	0	0	0
g	2	1	2	6	0	4	0	1	2	0	6	0	0	0	0	0	0	1	0	8	1	0	0	0	0	0
h	1	0	4	0	4	5	1	0	2	1	3	0	0	0	1	0	0	0	0	1	2	0	0	0	0	0
i	2	0	1	0	4	0	0	2	0	0	0	4	0	5	18	0	0	9	1	9	4	1	0	0	0	0
j	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	1	1	0	1	0	1	0	1	0	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0
l	3	1	1	1	2	0	0	0	6	0	0	0	0	0	4	9	0	0	0	1	2	0	0	0	1	0
m	1	0	1	0	0	0	0	1	1	0	0	2	0	4	3	0	0	0	1	0	0	0	0	0	0	0
n	2	2	3	1	4	0	3	0	0	0	0	0	4	0	22	1	0	2	3	2	6	0	1	0	0	0
o	3	1	5	0	4	0	1	1	11	0	0	7	3	1	0	12	0	6	1	2	1	0	0	0	0	0
p	0	0	0	0	0	0	0	0	0	0	0	3	0	1	16	0	0	0	2	0	0	2	0	0	0	0
q	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	5	0	2	1	19	2	0	0	7	0	0	6	0	1	3	2	0	0	3	13	7	1	0	0	1	0
s	9	1	17	9	6	0	0	2	3	0	0	1	1	2	2	0	0	2	0	6	0	0	5	1	0	0
t	4	0	2	5	3	1	4	8	2	0	0	1	2	6	0	0	0	8	2	0	2	0	0	0	4	0
u	0	0	0	0	0	0	1	1	0	0	0	1	0	1	8	0	0	8	0	0	0	0	4	0	11	0
v	0	12	8	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	2	0	0	0	0	0	0
w	0	0	0	0	14	0	0	0	0	0	0	0	0	0	2	0	0	1	1	2	0	0	0	0	0	0
x	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
y	2	0	0	1	0	0	0	5	0	0	0	1	0	0	0	1	0	4	0	9	0	0	0	0	0	0
z	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

表 A.2　入力すべき文字の前の文字の頻度を示した表（入力すべき文字が語頭の場合は_を示す）

誤り	入力すべき文字の前の文字の頻度																									
	_	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	14	3	1	2	1	5	2	0	0	0	0	0	1	4	3	0	4	0	9	1	1	1	0	2	0	0	0
b	4	1	0	0	1	1	0	0	0	4	0	0	1	0	0	0	0	0	1	0	0	2	0	0	0	0	0
c	4	3	0	0	0	4	0	0	0	4	0	0	0	0	0	2	0	0	3	0	0	0	0	0	0	0	0
d	7	3	0	0	1	5	0	0	0	4	0	0	0	0	3	2	0	0	6	1	1	1	0	0	0	0	0
e	8	2	0	2	6	8	4	3	2	4	0	3	5	3	2	5	2	0	14	11	5	4	0	0	0	0	0
f	5	0	0	0	1	2	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
g	2	7	0	1	0	1	1	0	0	6	0	0	1	0	11	2	0	0	1	0	0	0	0	1	0	0	0
h	7	0	0	6	0	0	1	2	1	0	0	0	0	1	0	0	0	0	3	0	1	2	0	1	0	0	0
i	9	4	0	4	2	1	0	1	5	3	0	0	2	3	2	1	4	1	7	3	3	3	2	0	0	0	0
j	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
k	3	8	0	1	0	0	0	0	0	2	0	0	0	0	0	1	0	0	3	0	0	1	0	1	0	0	0
l	2	4	1	2	0	3	2	0	1	1	0	0	2	6	0	4	0	0	1	0	1	1	0	0	0	0	0
m	4	2	0	0	0	1	0	0	0	0	0	0	0	2	0	2	0	0	2	0	1	0	0	0	0	0	0
n	5	3	0	1	0	1	0	0	1	24	0	0	1	1	3	7	1	0	1	1	0	6	0	0	0	0	0
o	19	2	1	6	1	1	1	0	0	4	0	0	3	1	0	5	2	0	3	3	6	0	0	1	0	0	0
p	5	0	0	0	0	3	1	0	0	0	1	0	5	0	0	4	0	0	0	0	5	0	0	0	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
r	7	4	1	3	0	12	2	0	10	2	0	0	5	0	0	6	3	0	5	0	6	3	1	3	0	0	0
s	21	7	0	0	0	7	0	0	0	1	0	0	6	0	5	6	2	0	3	4	0	3	0	2	0	0	0
t	4	2	0	0	0	7	0	9	2	4	0	0	1	0	4	6	0	0	3	2	4	6	0	0	0	0	0
u	1	2	1	2	0	0	0	1	3	1	0	0	2	0	0	8	3	0	7	2	2	0	0	0	0	0	0
v	9	2	0	1	0	2	0	0	0	6	0	0	0	0	1	1	0	0	0	0	0	3	0	0	0	0	0
w	8	0	0	0	1	2	0	0	0	0	1	0	1	0	0	0	5	0	0	0	0	0	1	1	0	0	0
x	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
y	0	4	1	2	0	0	0	1	0	3	0	0	0	0	1	1	1	0	2	3	3	1	0	0	0	0	0
z	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

表 A.3　入力すべき文字の後の文字の頻度を示した表（入力すべき文字が語末の場合は_を示す）

誤り	入力すべき文字の後の文字の頻度																									
	_	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	3	22	0	3	1	5	0	0	0	3	0	1	1	3	3	0	2	0	2	0	1	2	1	0	0	0	1
b	0	1	1	0	0	9	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
c	3	2	0	6	0	6	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
d	8	0	0	1	4	4	3	0	0	2	0	1	1	0	1	4	0	0	1	0	4	0	0	0	0	0	0
e	17	0	0	0	2	36	2	0	2	3	0	0	3	1	5	2	0	0	6	3	10	0	0	1	0	0	0
f	3	2	0	0	0	1	1	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0
g	17	2	0	0	0	3	1	7	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0
h	5	1	0	0	0	4	0	0	5	0	0	0	0	0	1	5	0	0	0	3	1	0	0	0	0	0	0
i	3	0	1	2	1	2	0	0	0	30	0	2	1	0	5	2	3	0	2	4	1	1	0	0	0	0	0
j	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	9	2	0	0	1	1	0	0	0	2	0	4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
l	4	0	0	1	0	1	0	0	0	0	0	0	21	0	0	1	0	0	1	1	0	0	0	1	0	0	0
m	2	2	0	1	0	1	0	0	0	0	0	0	1	3	0	2	0	0	1	1	0	0	0	0	0	0	0
n	5	1	1	0	0	1	0	0	2	0	0	0	1	1	40	2	0	0	0	0	1	1	0	0	0	0	0
o	4	0	0	1	1	4	0	1	0	6	0	0	5	1	2	24	1	0	4	3	0	0	1	1	0	0	0
p	4	0	1	2	0	2	1	0	0	0	0	0	0	0	0	2	5	0	1	0	0	1	1	4	0	0	0
q	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	19	2	0	1	0	7	2	1	1	0	0	1	1	0	5	1	0	0	26	0	3	2	0	0	0	0	0
s	9	3	0	4	0	15	2	0	6	7	0	0	0	0	3	2	0	0	3	10	1	0	0	0	0	2	0
t	16	1	0	1	0	1	0	2	2	0	0	0	0	0	3	2	0	0	1	2	21	2	0	0	0	0	0
u	10	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	2	6	14	0	0	0	0	0
v	2	2	0	1	0	11	0	0	0	1	0	0	6	0	0	0	0	0	1	0	1	0	0	0	0	0	0
w	2	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	3	4	0	1	1	0	5	2	0	0
x	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
y	9	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	2	0	0	0	0	0	0	9	0
z	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0